On the Maximal Sum of Exponents of Runs in a String

Maxime Crochemore, Marcin Kubica, Jakub Radoszewski, Wojciech Rytter, Tomasz Walen

On the Maximal Sum of Exponents of Runs in a String

Maxime Crochemore[1,3], Marcin Kubica[2], Jakub Radoszewski[⋆2], Wojciech Rytter[2,5], and Tomasz Waleń[2]

[1] King's College London, London WC2R 2LS, UK
maxime.crochemore@kcl.ac.uk
[2] Dept. of Mathematics, Computer Science and Mechanics, University of Warsaw, Warsaw, Poland
[kubica,jrad,rytter,walen]@mimuw.edu.pl
[3] Université Paris-Est, France
[4] Dept. of Math. and Informatics, Copernicus University, Toruń, Poland

Abstract. A run is an inclusion maximal occurrence in a string (as a subinterval) of a repetition v with a period p such that $2p \leq |v|$. The exponent of a run is defined as $|v|/p$ and is ≥ 2. We show new bounds on the maximal sum of exponents of runs in a string of length n. Our upper bound of $4.1\,n$ is better than the best previously known proven bound of $5.6\,n$ by Crochemore & Ilie (2008). The lower bound of $2.035\,n$, obtained using a family of binary words, contradicts the conjecture of Kolpakov & Kucherov (1999) that the maximal sum of exponents of runs in a string of length n is smaller than $2n$.

1 Introduction

Repetitions and periodicities in strings are one of the fundamental topics in combinatorics on words [1, 14]. They are also important in other areas: lossless compression, word representation, computational biology, etc. In this paper we consider bounds on the sum of exponents of repetitions that a string of a given length may contain. In general, repetitions are studied also from other points of view, like: the classification of words (both finite and infinite) not containing repetitions of a given exponent, efficient identification of factors being repetitions of different types and computing the bounds on the number of various types of repetitions occurring in a string. The known results in the topic and a deeper description of the motivation can be found in a survey by Crochemore et al. [4].

The concept of runs (also called maximal repetitions) has been introduced to represent all repetitions in a string in a succinct manner. The crucial property of runs is that their maximal number in a string of length n (denoted as $\rho(n)$) is $O(n)$, see Kolpakov & Kucherov [10]. This fact is the cornerstone of any

⋆ Some parts of this paper were written during the author's Erasmus exchange at King's College London

algorithm computing all repetitions in strings of length n in $O(n)$ time. Due to the work of many people, much better bounds on $\rho(n)$ have been obtained. The lower bound $0.927\,n$ was first proved by Franek & Yang [7]. Afterwards, it was improved by Kusano et al. [13] to $0.944565\,n$ employing computer experiments, and very recently by Simpson [18] to $0.944575712\,n$. On the other hand, the first explicit upper bound $5\,n$ was settled by Rytter [16], afterwards it was systematically improved to $3.48\,n$ by Puglisi et al. [15], $3.44\,n$ by Rytter [17], $1.6\,n$ by Crochemore & Ilie [2, 3] and $1.52\,n$ by Giraud [8]. The best known result $\rho(n) \leq 1.029\,n$ is due to Crochemore et al. [5], but it is conjectured [10] that $\rho(n) < n$. Some results are known also for repetitions of exponent higher than 2. For instance, the maximal number of cubic runs (maximal repetitions with exponent at least 3) in a string of length n (denoted $\rho_{cubic}(n)$) is known to be between $0.406\,n$ and $0.5\,n$, see Crochemore et al. [6].

A stronger property of runs is that the maximal sum of their exponents in a string of length n (notation: $\sigma(n)$) is linear in terms of n, see Kolpakov & Kucherov [12]. It has applications to the analysis of various algorithms, such as computing branching tandem repeats: the linearity of the sum of exponents solves a conjecture of [9] concerning the linearity of the number of maximal tandem repeats and implies that all can be found in linear time. For other applications, we refer to [12]. The proof that $\sigma(n) < cn$ in Kolpakov and Kucherov's paper [12] is very complex and does not provide any particular value for the constant c. A bound can be derived from the proof of Rytter [16] but he mentioned only that the bound that he obtains is "unsatisfactory" (it seems to be $25\,n$). The first explicit bound $5.6\,n$ for $\sigma(n)$ was provided by Crochemore and Ilie [3], who claim that it could be improved to $2.9\,n$ employing computer experiments. As for the lower bound on $\sigma(n)$, no exact values were previously known and it was conjectured [11, 12] that $\sigma(n) < 2n$.

In this paper we provide an upper bound of $4.1\,n$ on the maximal sum of exponents of runs in a string of length n and also a stronger upper bound of $2.5\,n$ for the maximal sum of exponents of cubic runs in a string of length n. As for the lower bound, we bring down the conjecture $\sigma(n) < 2n$ by providing an infinite family of binary strings for which the sum of exponents of runs is greater than $2.035\,n$.

2 Preliminaries

We consider *words* (*strings*) u over a finite alphabet Σ, $u \in \Sigma^*$; the empty word is denoted by ε; the positions in u are numbered from 1 to $|u|$. For $u = u_1u_2\ldots u_m$, let us denote by $u[i\,..\,j]$ a *factor* of u equal to $u_i \ldots u_j$ (in particular $u[i] = u[i\,..\,i]$). Words $u[1\,..\,i]$ are called prefixes of u, and words $u[i\,..\,|u|]$ suffixes of u.

We say that an integer p is the (shortest) *period* of a word $u = u_1 \ldots u_m$ (notation: $p = \mathsf{per}(u)$) if p is the smallest positive integer such that $u_i = u_{i+p}$ holds for all $1 \leq i \leq m - p$. We say that words u and v are cyclically equivalent

(or that one of them is a cyclic rotation of the other) if $u = xy$ and $v = yx$ for some $x, y \in \Sigma^*$.

A *run* (also called a maximal repetition) in a string u is an interval $[i\,..\,j]$ such that:

- the period p of the associated factor $u[i\,..\,j]$ satisfies $2p \leq j - i + 1$,
- the interval cannot be extended to the right nor to the left, without violating the above property, that is, $u[i-1] \neq u[i+p-1]$ and $u[j-p+1] \neq u[j+1]$.

A *cubic run* is a run $[i\,..\,j]$ for which the shortest period p satisfies $3p \leq j-i+1$. For simplicity, in the rest of the text we sometimes refer to runs and cubic runs as to occurrences of the corresponding factors of u. The (fractional) *exponent* of a run is defined as $(j - i + 1)/p$.

For a given word $u \in \Sigma^*$, we introduce the following notation:

- $\rho(u)$ and $\rho_{cubic}(u)$ are the numbers of runs and cubic runs in u resp.
- $\sigma(u)$ and $\sigma_{cubic}(u)$ are the sums of exponents of runs and cubic runs in u resp.

For a non-negative integer n, we use the same notations $\rho(n)$, $\rho_{cubic}(n)$, $\sigma(n)$ and $\sigma_{cubic}(n)$ to denote the maximal value of the respective function for a word of length n.

3 Lower bound for $\sigma(n)$

Tables 1 and 2 list the sums of exponents of runs for several words of two known families that contain very large number of runs: the words x_i defined by Franek and Yang [7] (giving the lower bound $\rho(n) \geq 0.927\,n$, conjectured for some time to be optimal) and the modified Padovan words y_i defined by Simpson [18] (giving the best known lower bound $\rho(n) \geq 0.944575712\,n$). These values have been computed experimentally. They suggest that for the families of words x_i and y_i the maximal sum of exponents could be less than $2n$.

We show, however, a lower bound for $\sigma(n)$ that is greater than $2n$.

Theorem 1. *There are infinitely many binary strings w such that*

$$\frac{\sigma(w)}{|w|} > 2.035.$$

Proof. Let us define two morphisms $\phi : \{a, b, c\} \mapsto \{a, b, c\}$ and $\psi : \{a, b, c\} \mapsto \{0, 1\}$ as follows:

$$\phi(a) = baaba, \quad \phi(b) = ca, \quad \phi(c) = bca$$

$$\psi(a) = 01011, \quad \psi(b) = \psi(c) = 01001011$$

We define $w_i = \psi(\phi^i(a))$. Table 3 shows the sums of exponents of runs in words w_i, computed experimentally.

Clearly, for any word $w = (w_8)^k$, $k \geq 1$, we have

$$\frac{\sigma(w)}{|w|} > 2.035.$$

□

i	$\lvert x_i\rvert$	$\rho(x_i)/\lvert x_i\rvert$	$\sigma(x_i)$	$\sigma(x_i)/\lvert x_i\rvert$
1	6	0.3333	4.00	0.6667
2	27	0.7037	39.18	1.4510
3	116	0.8534	209.70	1.8078
4	493	0.9047	954.27	1.9356
5	2090	0.9206	4130.66	1.9764
6	8855	0.9252	17608.48	1.9885
7	37512	0.9266	74723.85	1.9920
8	158905	0.9269	316690.85	1.9930
9	673134	0.9270	1341701.95	1.9932

Table 1. Number of runs and sum of exponents of runs in Franek & Yang's [7] words x_i.

i	$\lvert y_i\rvert$	$\rho(y_i)/\lvert y_i\rvert$	$\sigma(y_i)$	$\sigma(y_i)/\lvert y_i\rvert$
4	37	0.7568	57.98	1.5671
8	125	0.8640	225.75	1.8060
12	380	0.9079	726.66	1.9123
16	1172	0.9309	2303.21	1.9652
20	3609	0.9396	7165.93	1.9856
24	11114	0.9427	22148.78	1.9929
28	34227	0.9439	68307.62	1.9957
32	105405	0.9443	210467.18	1.9967
36	324605	0.9445	648270.74	1.9971
40	999652	0.9445	1996544.30	1.9972

Table 2. Number of runs and sum of exponents of runs in Simpson's [18] modified Padovan words y_i.

i	$\lvert w_i\rvert$	$\sigma(w_i)$	$\sigma(w_i)/\lvert w_i\rvert$
1	31	47.10	1.5194
2	119	222.26	1.8677
3	461	911.68	1.9776
4	1751	3533.34	2.0179
5	6647	13498.20	2.0307
6	25205	51264.37	2.0339
7	95567	194470.30	2.0349
8	362327	737393.11	2.0352
9	1373693	2795792.39	2.0352
10	5208071	10599765.15	2.0353

Table 3. Sums of exponents of runs in words w_i.

4 Upper bounds for $\sigma(n)$ and $\sigma_{cubic}(n)$

In this section we utilize the concept of *handles* of runs as defined in [6]. The original definition refers only to cubic runs, but here we extend it also to ordinary runs.

Let $u \in \Sigma^*$ be a word of length n. Let us denote by $P = \{p_1, p_2, \ldots, p_{n-1}\}$ the set of inter-positions in u that are located *between* pairs of consecutive letters of u. We define a function H assigning to each run v in u a set of some inter-positions within v (called later on *handles*) — H is a mapping from the set of runs occurring in u to the set 2^P of subsets of P. Let v be a run with period p and let w be the prefix of v of length p. Let $w_{\min}$ and $w_{\max}$ be the minimal and maximal words (in lexicographical order) cyclically equivalent to w. $H(v)$ is defined as follows:

a) if $w_{\min} = w_{\max}$ then $H(v)$ contains all inter-positions within v,
b) if $w_{\min} \neq w_{\max}$ then $H(v)$ contains inter-positions between consecutive occurrences of $w_{\min}$ in v and between consecutive occurrences of $w_{\max}$ in v.

Note that $H(v)$ can be empty for a non-cubic-run v.

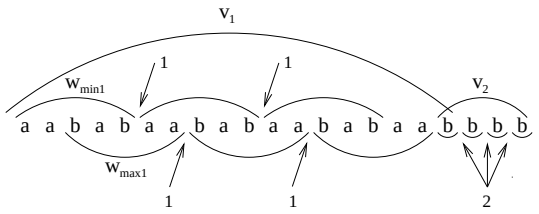

Fig. 1. An example of a word with two highlighted runs v_1 and v_2. For v_1 we have $w_{\min 1} \neq w_{\max 1}$ and for v_2 the corresponding words are equal to b (a one-letter word). The inter-positions belonging to the sets $H(v_1)$ and $H(v_2)$ are pointed by arrows

Proofs of the following properties of handles of runs can be found in [6]:

1. Case (a) in the definition of $H(v)$ implies that $|w_{\min}| = 1$.
2. $H(v_1) \cap H(v_2) = \emptyset$ for any two distinct runs v_1 and v_2 in u.

To prove the upper bound for $\sigma(n)$, we need to state an additional property of handles of runs. Let $\mathcal{R}(u)$ be the set of all runs in a word u, and let $\mathcal{R}_1(u)$ and $\mathcal{R}_{\geq 2}(u)$ be the sets of runs with period 1 and at least 2 respectively.

Lemma 1.
If $v \in \mathcal{R}_1(u)$ *then* $\sigma(v) = |H(v)| + 1$.
If $v \in \mathcal{R}_{\geq 2}(u)$ *then* $\lceil \sigma(v) \rceil \leq \frac{|H(v)|}{2} + 3$.

Proof. For the case of $v \in \mathcal{R}_1(u)$, the proof is straightforward from the definition of handles. In the opposite case, it is sufficient to note that both words $w^k_{\min}$ and $w^k_{\max}$ for $k = \lfloor \sigma(v) \rfloor - 1$ are factors of v, and thus

$$|H(v)| \geq 2 \cdot (\lfloor \sigma(v) \rfloor - 2).$$

□

Now we are ready to prove the upper bound for $\sigma(n)$. In the proof we use the bound $\rho(n) \leq 1.029\, n$ on the number of runs from [5].

Theorem 2. *The sum of the exponents of runs in a string of length n is less than $4.1\, n$.*

Proof. Let u be a word of length n. Using Lemma 1, we obtain:

$$\begin{aligned}
\sum_{v \in \mathcal{R}(u)} \sigma(v) &= \sum_{v \in \mathcal{R}_1(u)} \sigma(v) + \sum_{v \in \mathcal{R}_{\geq 2}(u)} \sigma(v) \\
&\leq \sum_{v \in \mathcal{R}_1(u)} (|H(v)| + 1) + \sum_{v \in \mathcal{R}_{\geq 2}(u)} \left(\frac{|H(v)|}{2} + 3 \right) \\
&= \sum_{v \in \mathcal{R}_1(u)} |H(v)| + |\mathcal{R}_1(u)| + \sum_{v \in \mathcal{R}_{\geq 2}(u)} \frac{|H(v)|}{2} + 3 \cdot |\mathcal{R}_{\geq 2}(u)| \\
&\leq 3 \cdot |\mathcal{R}(u)| + A + B/2, \qquad (1)
\end{aligned}$$

where $A = \sum_{v \in \mathcal{R}_1(u)} |H(v)|$ and $B = \sum_{v \in \mathcal{R}_{\geq 2}(u)} |H(v)|$. Due to the disjointness of handles of runs (the second property of handles), $A + B < n$, and thus, $A + B/2 < n$. Combining this with (1), we obtain:

$$\sum_{v \in \mathcal{R}(u)} \sigma(v) \;<\; 3 \cdot |\mathcal{R}(u)| + n \;\leq\; 3 \cdot \rho(n) + n \;\leq\; 3 \cdot 1.029\, n + n \;<\; 4.1\, n.$$

□

A similar approach for cubic runs, this time using the bound of $0.5\, n$ for $\rho_{cubic}(n)$ from [6], enables us to immediately provide a stronger upper bound for the function $\sigma_{cubic}(n)$.

Theorem 3. *The sum of the exponents of cubic runs in a string of length n is less than $2.5\, n$.*

Proof. Let u be a word of length n. Using same inequalities as in the proof of Theorem 2, we obtain:

$$\sum_{v \in \mathcal{R}_{cubic}(u)} \sigma(v) \;<\; 3 \cdot |\mathcal{R}_{cubic}(u)| + n \;\leq\; 3 \cdot \rho_{cubic}(n) + n \;\leq\; 3 \cdot 0.5\, n + n \;=\; 2.5\, n,$$

where $\mathcal{R}_{cubic}(u)$ denotes the set of all cubic runs of u. □

References

1. J. Berstel and J. Karhumaki. Combinatorics on words: a tutorial. *Bulletin of the EATCS*, 79:178–228, 2003.
2. M. Crochemore and L. Ilie. Analysis of maximal repetitions in strings. In L. Kucera and A. Kucera, editors, *MFCS*, volume 4708 of *Lecture Notes in Computer Science*, pages 465–476. Springer, 2007.
3. M. Crochemore and L. Ilie. Maximal repetitions in strings. *J. Comput. Syst. Sci.*, 74(5):796–807, 2008.
4. M. Crochemore, L. Ilie, and W. Rytter. Repetitions in strings: Algorithms and combinatorics. *Theor. Comput. Sci.*, 410(50):5227–5235, 2009.
5. M. Crochemore, L. Ilie, and L. Tinta. Towards a solution to the "runs" conjecture. In P. Ferragina and G. M. Landau, editors, *CPM*, volume 5029 of *Lecture Notes in Computer Science*, pages 290–302. Springer, 2008.
6. M. Crochemore, C. Iliopoulos, M. Kubica, J. Radoszewski, W. Rytter, and T. Walen. On the maximal number of cubic runs in a string. In *Proceedings of LATA*, 2010 (to appear).
7. F. Franek and Q. Yang. An asymptotic lower bound for the maximal number of runs in a string. *Int. J. Found. Comput. Sci.*, 19(1):195–203, 2008.
8. M. Giraud. Not so many runs in strings. In C. Martín-Vide, F. Otto, and H. Fernau, editors, *LATA*, volume 5196 of *Lecture Notes in Computer Science*, pages 232–239. Springer, 2008.
9. D. Gusfield and J. Stoye. Simple and flexible detection of contiguous repeats using a suffix tree (preliminary version). In M. Farach-Colton, editor, *CPM*, volume 1448 of *Lecture Notes in Computer Science*, pages 140–152. Springer, 1998.
10. R. M. Kolpakov and G. Kucherov. Finding maximal repetitions in a word in linear time. In *Proceedings of the 40th Symposium on Foundations of Computer Science*, pages 596–604, 1999.
11. R. M. Kolpakov and G. Kucherov. On maximal repetitions in words. *J. of Discr. Alg.*, 1:159–186, 1999.
12. R. M. Kolpakov and G. Kucherov. On the sum of exponents of maximal repetitions in a word. *Tech. Report 99-R-034, LORIA*, 1999.
13. K. Kusano, W. Matsubara, A. Ishino, H. Bannai, and A. Shinohara. New lower bounds for the maximum number of runs in a string. *CoRR*, abs/0804.1214, 2008.
14. M. Lothaire. *Combinatorics on Words.* Addison-Wesley, Reading, MA., U.S.A., 1983.
15. S. J. Puglisi, J. Simpson, and W. F. Smyth. How many runs can a string contain? *Theor. Comput. Sci.*, 401(1-3):165–171, 2008.
16. W. Rytter. The number of runs in a string: Improved analysis of the linear upper bound. In B. Durand and W. Thomas, editors, *STACS*, volume 3884 of *Lecture Notes in Computer Science*, pages 184–195. Springer, 2006.
17. W. Rytter. The number of runs in a string. *Inf. Comput.*, 205(9):1459–1469, 2007.
18. J. Simpson. Modified Padovan words and the maximum number of runs in a word. *Australasian J. of Comb.*, 46:129–145, 2010.